t:@rmoff e:robin@confluent.io

Apache Kafka™'s Role in Implementing Oracle's Big Data Reference Architecture

SUN6259

Oracle OpenWorld, 1 Oct 2017

Robin Moffatt, Partner Technology Evangelist, EMEA

What is Apache Kafka™?

“Apache Kafka™ is a distributed streaming platform”

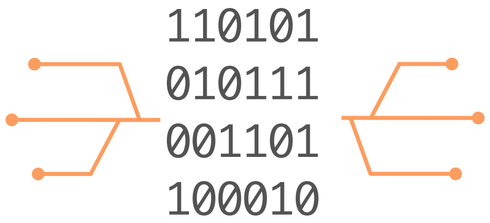
Sounds Fancy

But

What *is* Kafka?

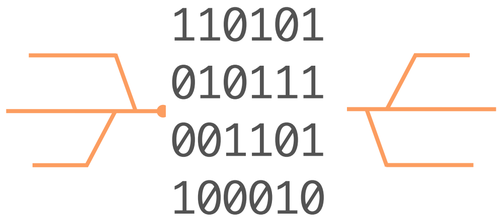
Kafka is a Distributed Streaming Platform

Publish and subscribe to streams of data



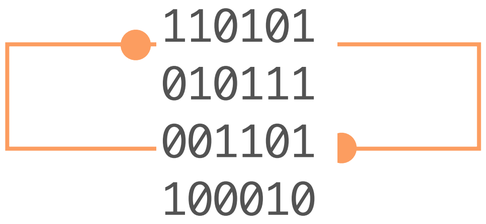
similar to a message queue or enterprise messaging system.

Store streams of data



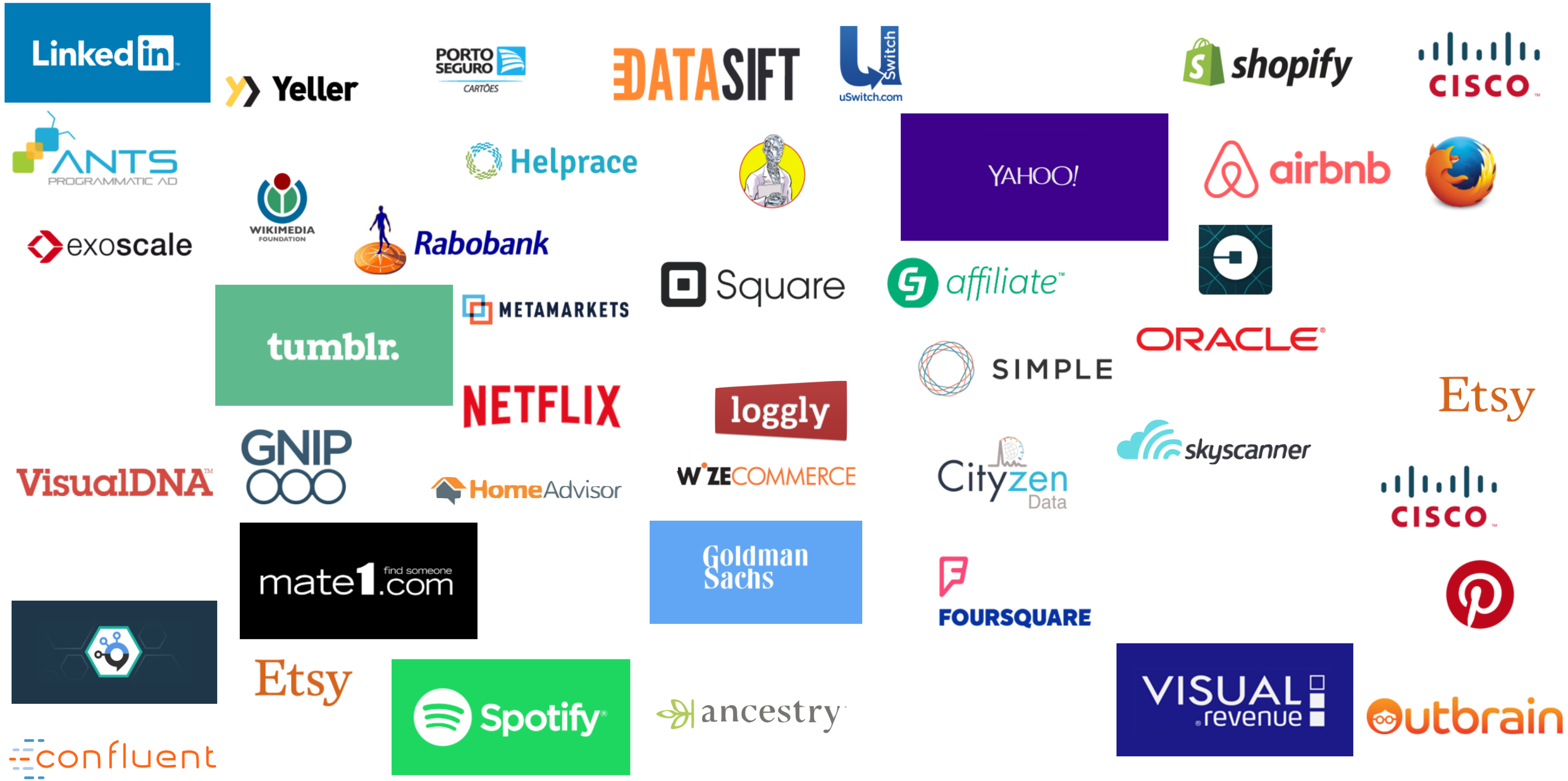
in a fault tolerant way.

Process streams of data



in real time, as they occur.

Powered by Apache Kafka™



$ whoami

- Partner Technology Evangelist @ Confluent
- Working in data & analytics since 2001
- Oracle ACE Director
- Blogging : http://rmoff.net & https://www.confluent.io/blog/author/robin/
- Twitter: @rmoff
  - Geek stuff
  - Beer & Fried Breakfasts

confluent

The Reference Architecture

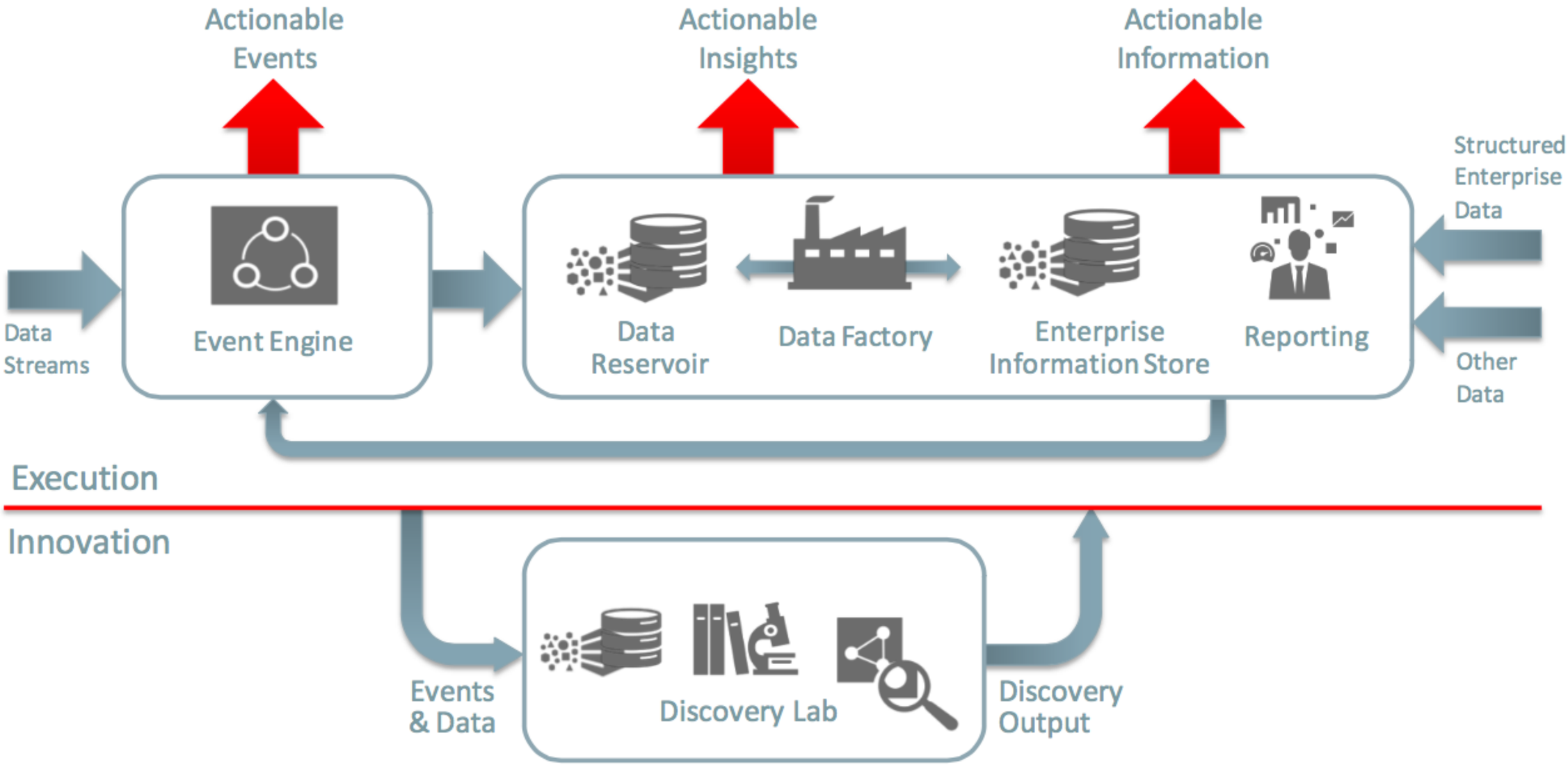
Or “Information Management and Big Data - A Reference Architecture” for short…

Information Management and Big Data Reference Architecture

- Tool-agnostic logical architecture for Information Management, taking into account Big Data
- Written by Oracle, with input from Mark Rittman and Stewart Bryson
- Three years old, but still a good starting point for implementation design

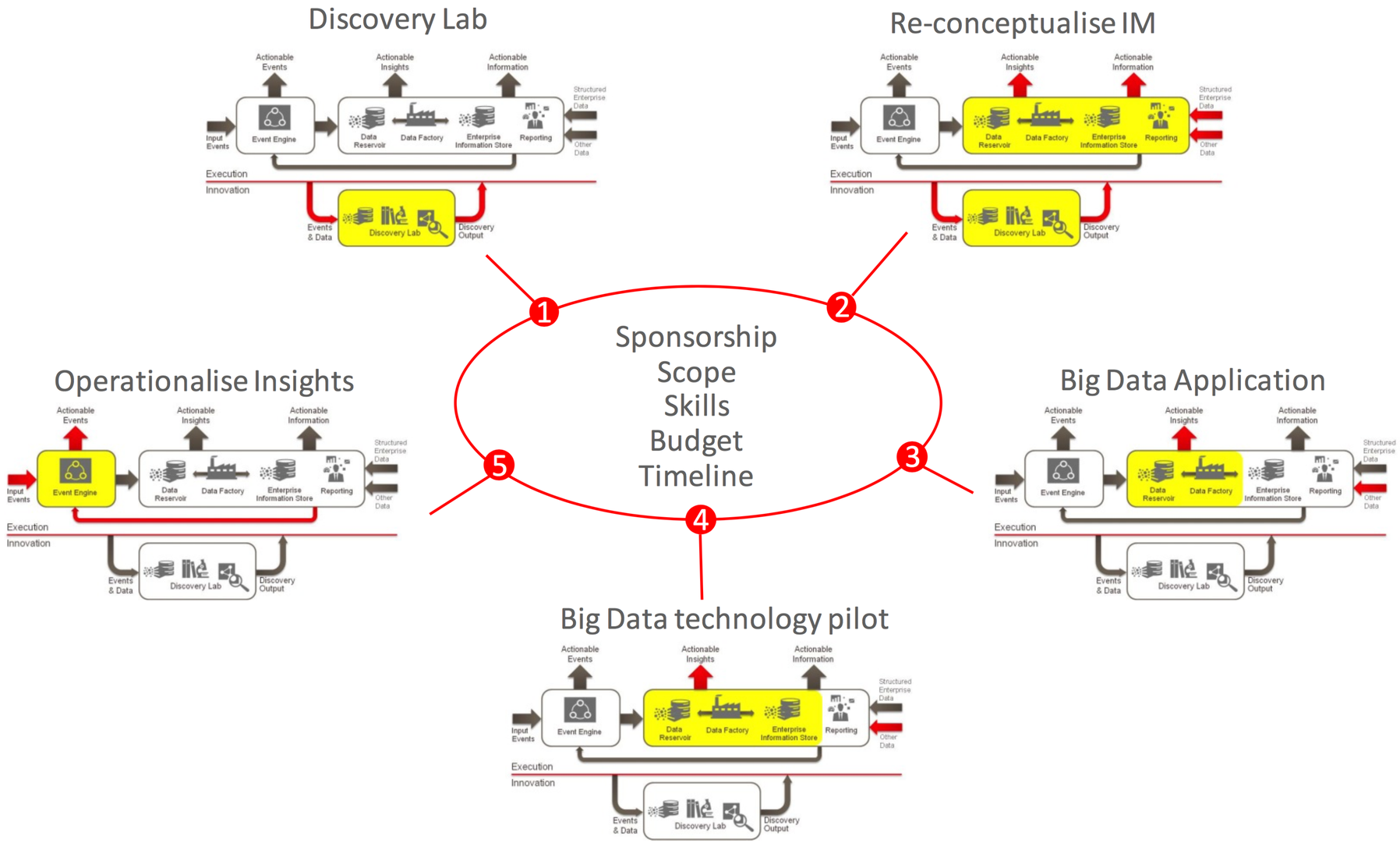
http://www.oracle.com/technetwork/database/bigdata-appliance/overview/bigdatarefarchitecture-2297765.pdf

Conceptual Architecture



http://www.oracle.com/technetwork/database/bigdata-appliance/overview/bigdatarefarchitecture-2297765.pdf

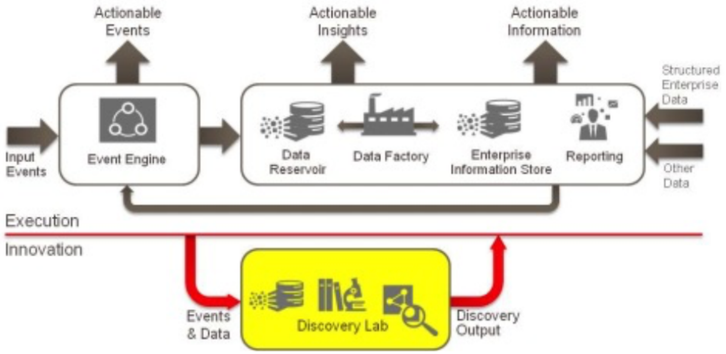
Implementation Patterns



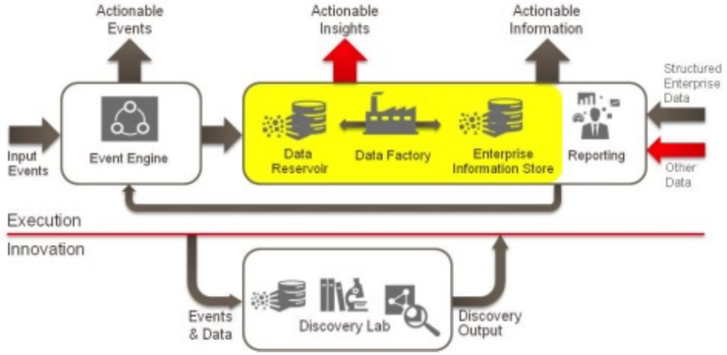
http://www.oracle.com/technetwork/database/bigdata-appliance/overview/bigdatarefarchitecture-2297765.pdf

How Do We Build For the Future?

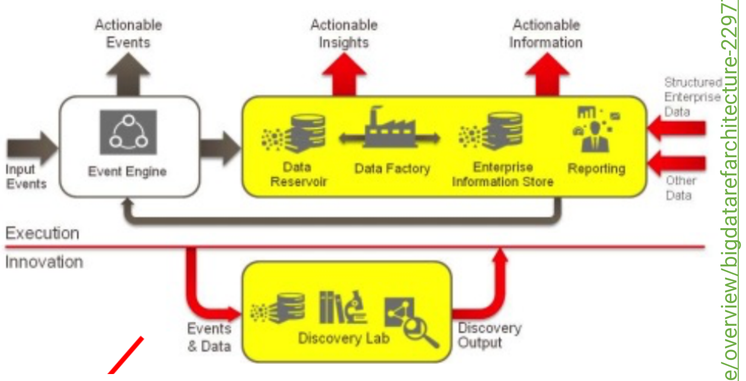
Discovery Lab



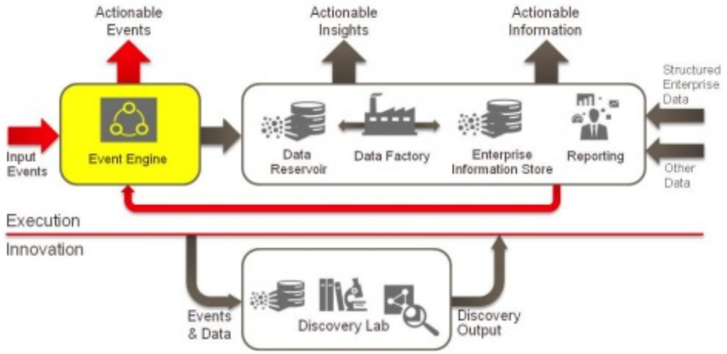
Big Data technology pilot



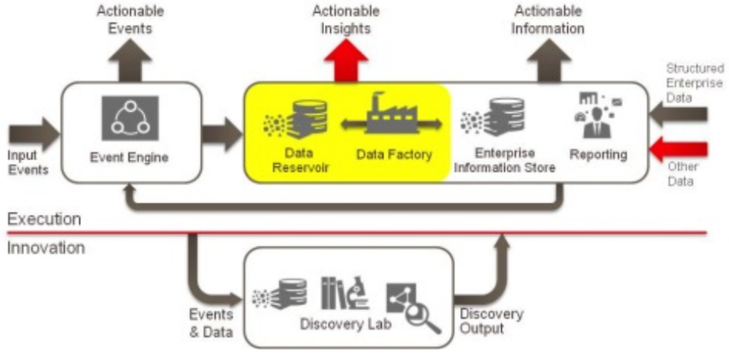
Re-conceptualise IM



Operationalise Insights



Big Data Application



http://www.oracle.com/technetwork/database/bigdata-appliance/overview/bigdatarefarchitecture-2297765.pdf

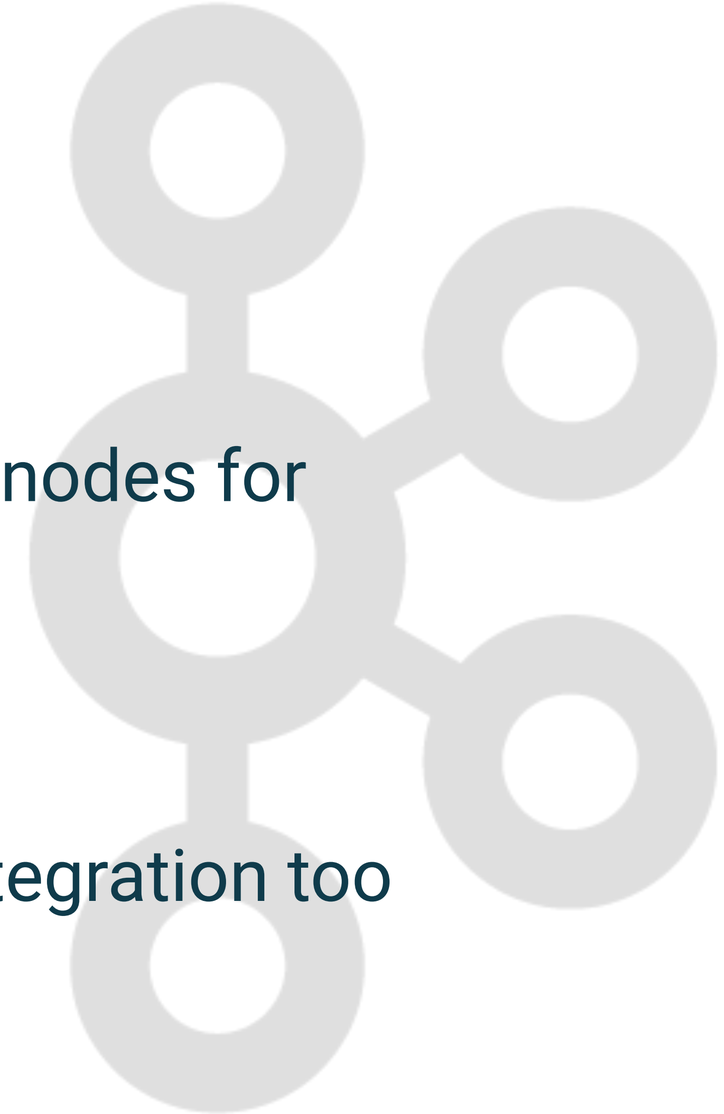
Kafka in Detail

What is Kafka?

- Messages are stored in Topics
  - Roughly analogous to a database table

- Topics can be partitioned across multiple Kafka nodes for redundancy and performance

- Not just about streaming - huge uses for data integration too



What is Kafka?

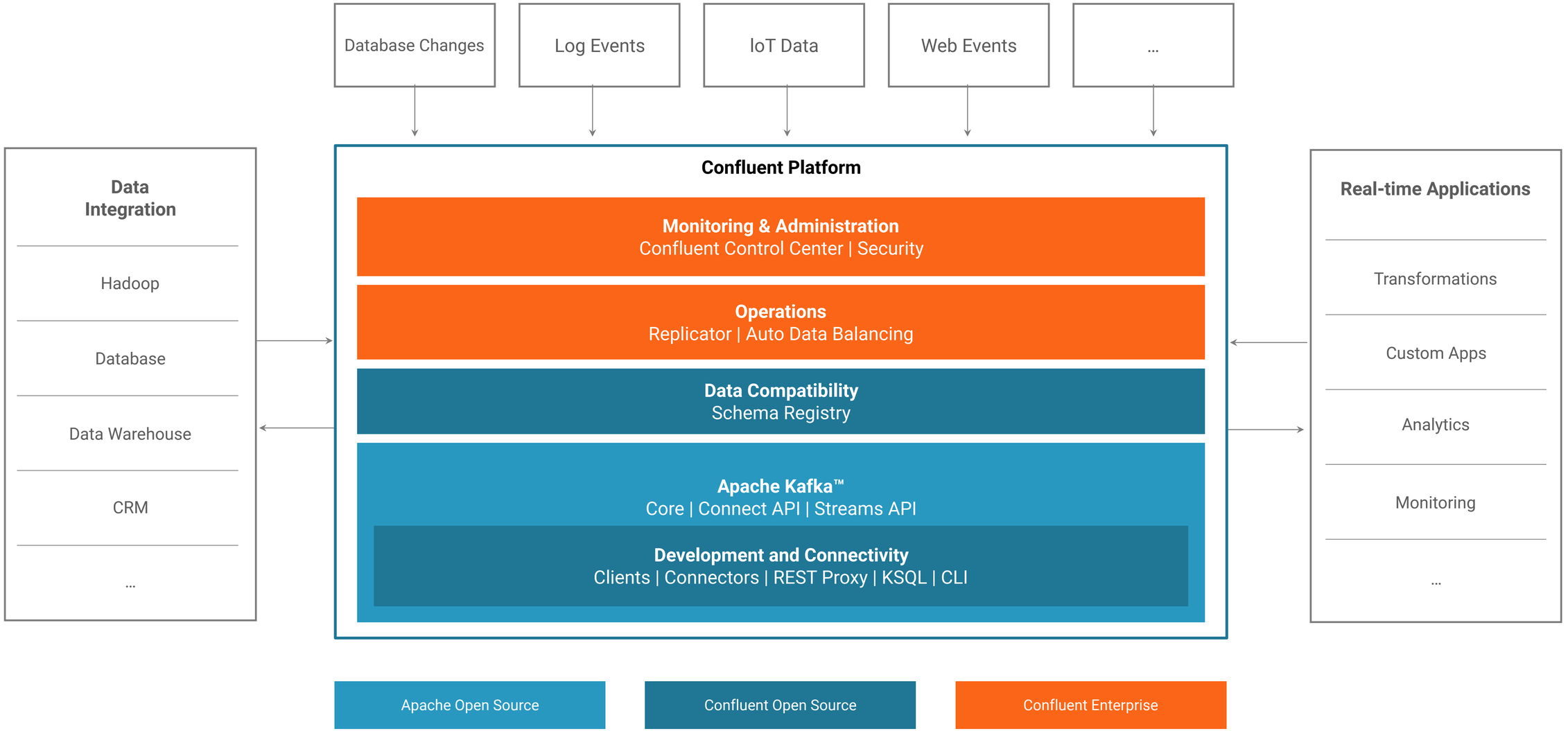
- Kafka makes its data available to any consumer
  - Security permitting

- Consumers: -
  - Are independent from each other
  - Can be grouped and parallelised for performance and resilience
  - Can re-read messages as required
  - Can read messages stream or batch

Running Kafka

- Apache Kafka is open source
  - Includes Kafka Core, streams processing and data integration capabilities
- Can be deployed standalone or as part of Confluent Platform
- Also available in most Hadoop distributions, but older versions without latest functionality

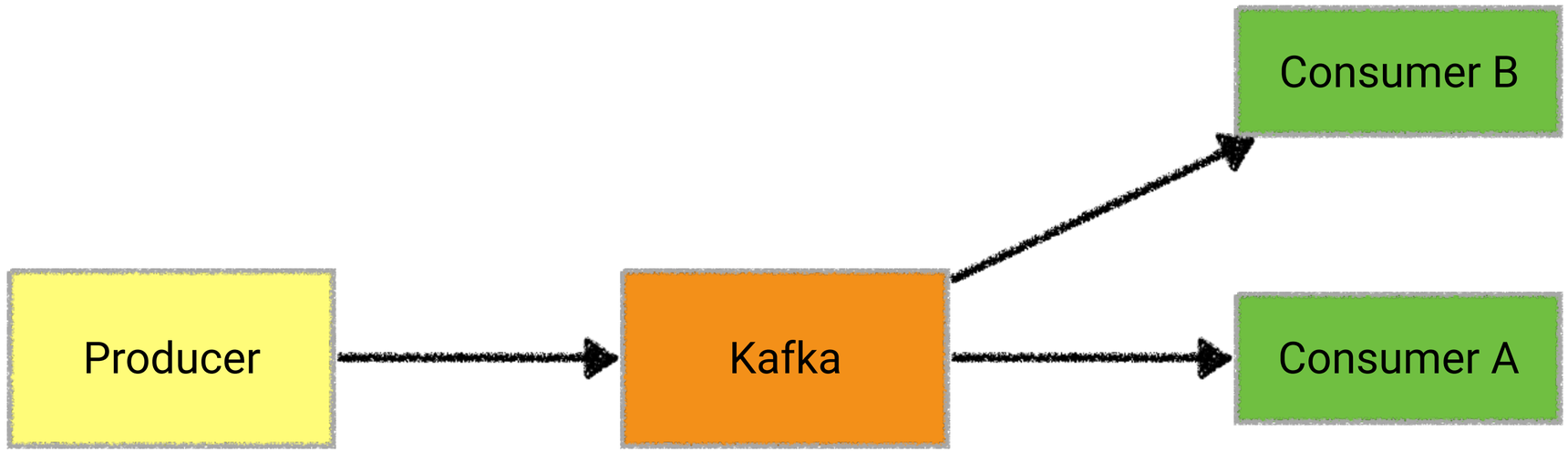
Confluent Platform: Enterprise Streaming based on Apache Kafka™



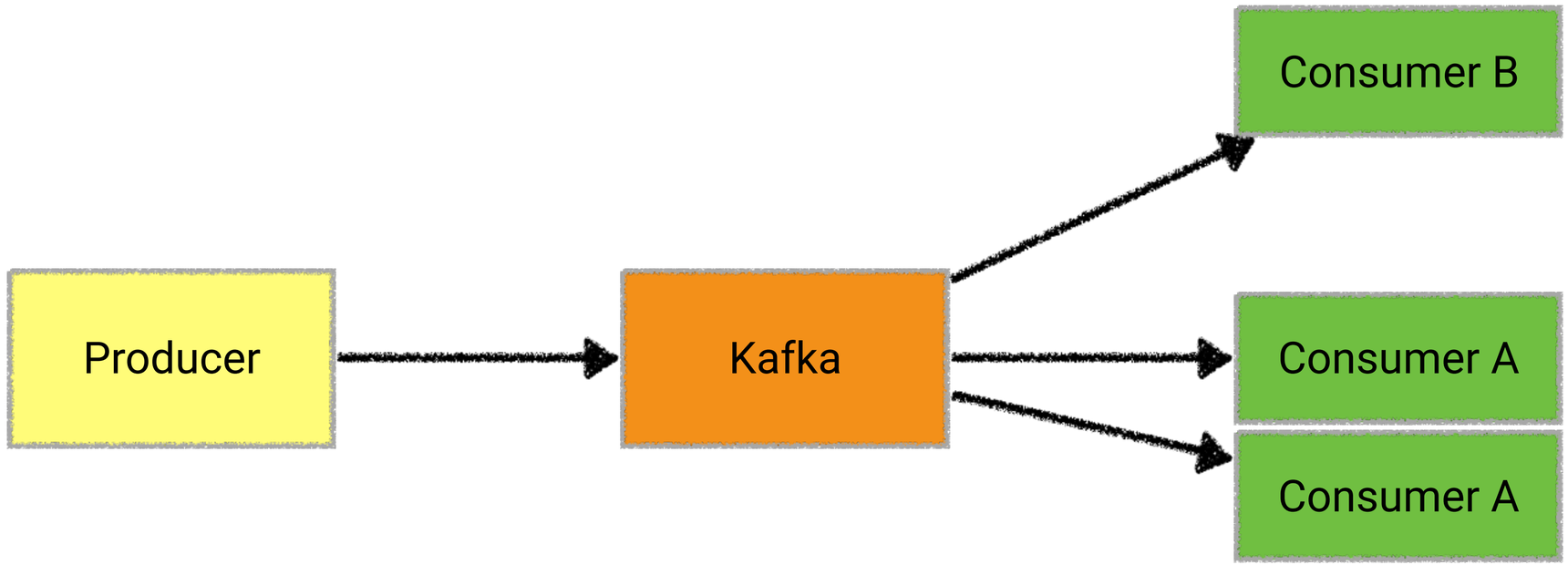
What are the Problems that Kafka Solves?

What are the Problems That Kafka Solves?

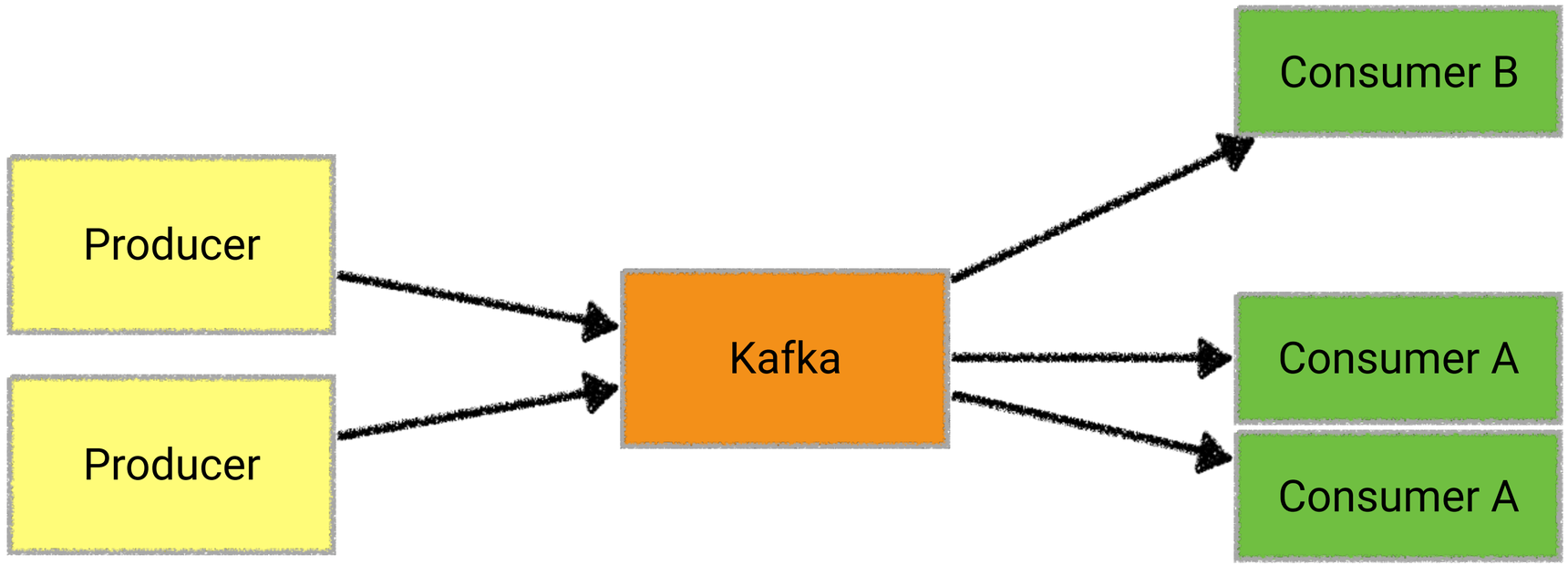
Multiple Independent Customers of the Same Data



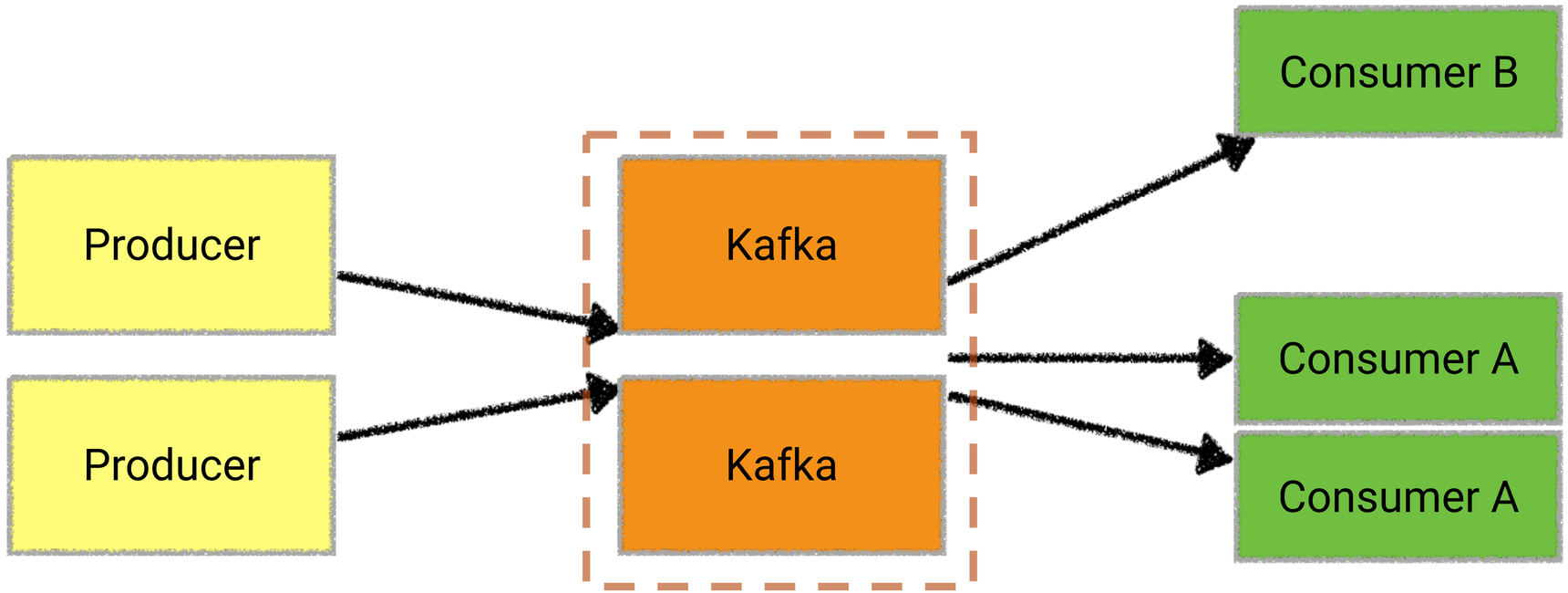
Multiple Parallel Customers of the Same Data



Multiple Sources of the Same Type of Data



Scaling Throughput and Resilience

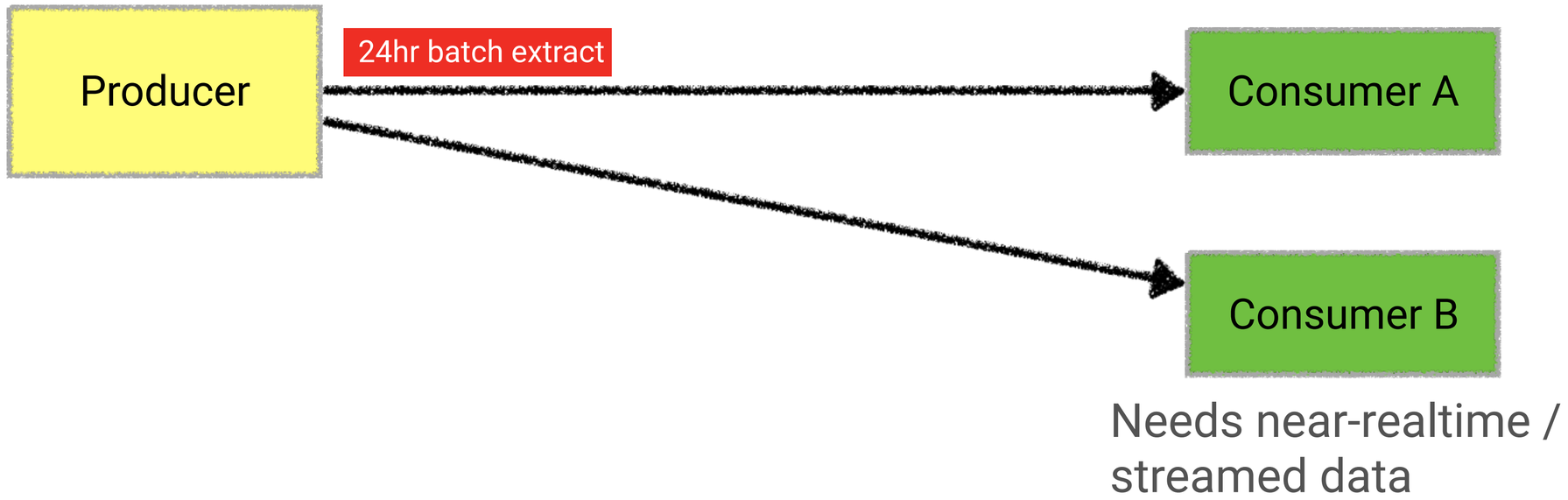


System Availability and Event Buffering

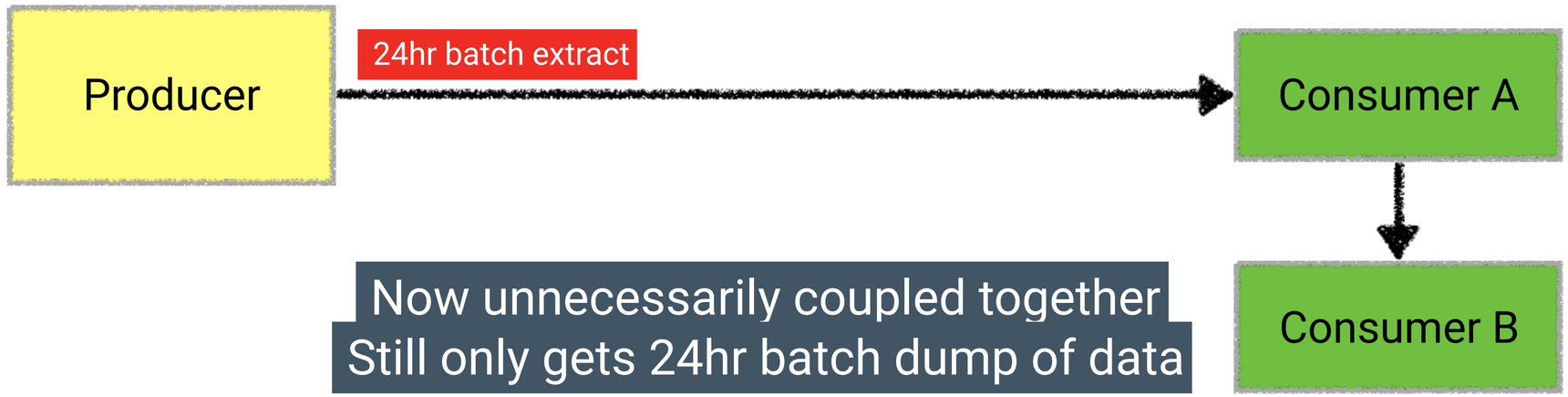
System Availability and Event Buffering

Varying Latency Requirements / Batch vs Stream

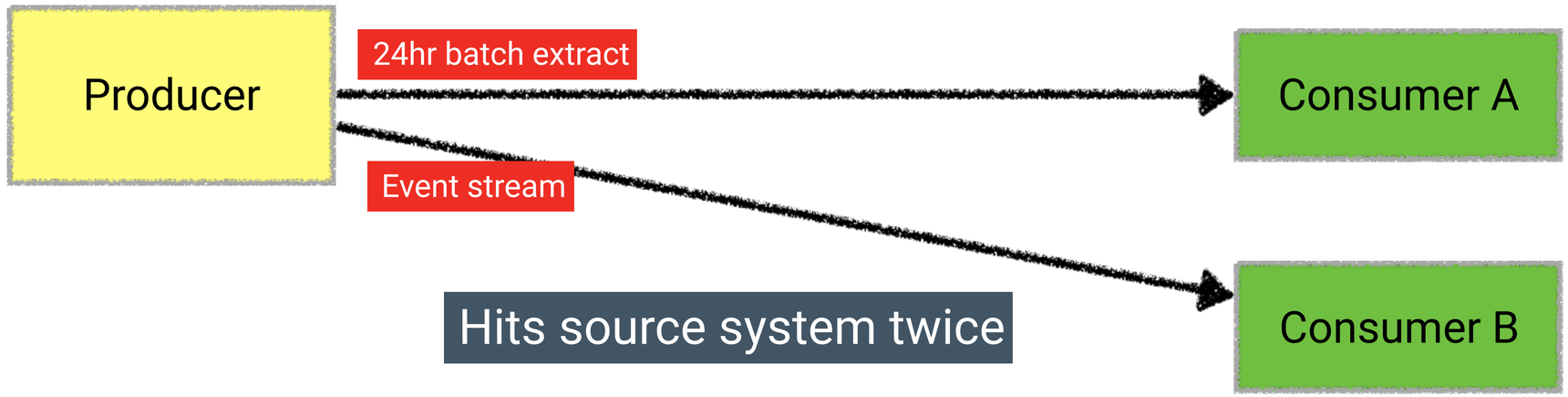
Varying Latency Requirements / Batch vs Stream



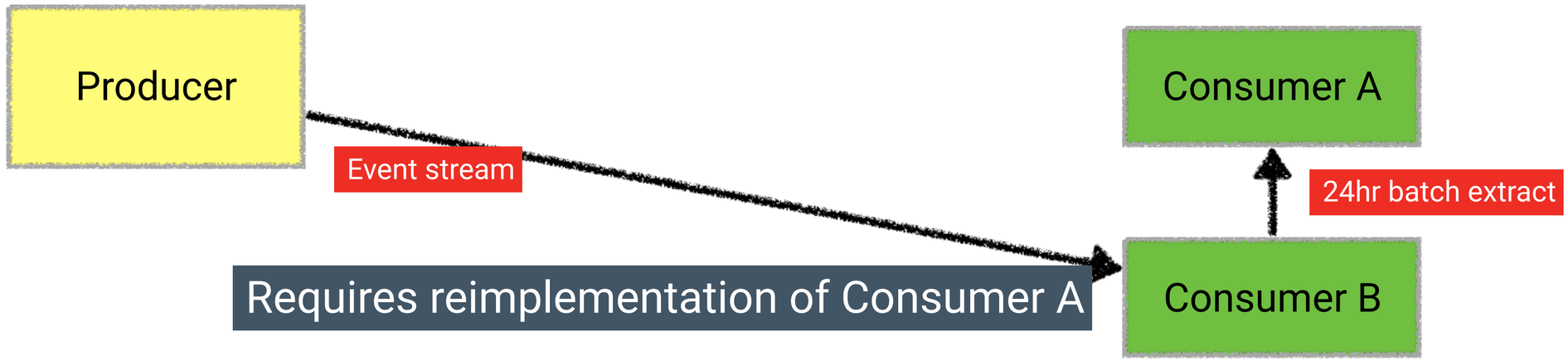
Varying Latency Requirements / Batch vs Stream



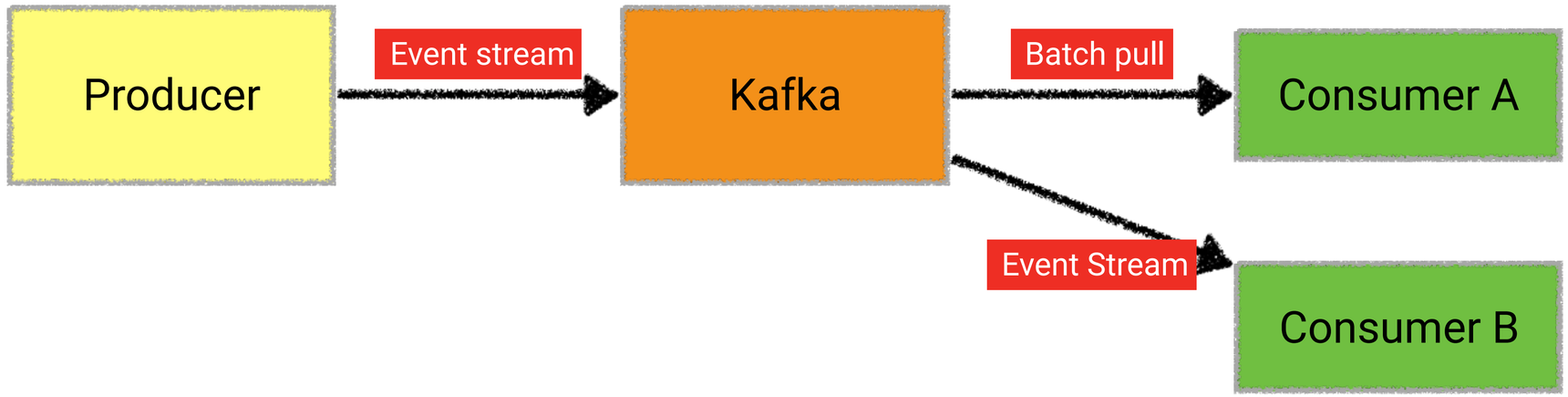
Varying Latency Requirements / Batch vs Stream



Varying Latency Requirements / Batch vs Stream

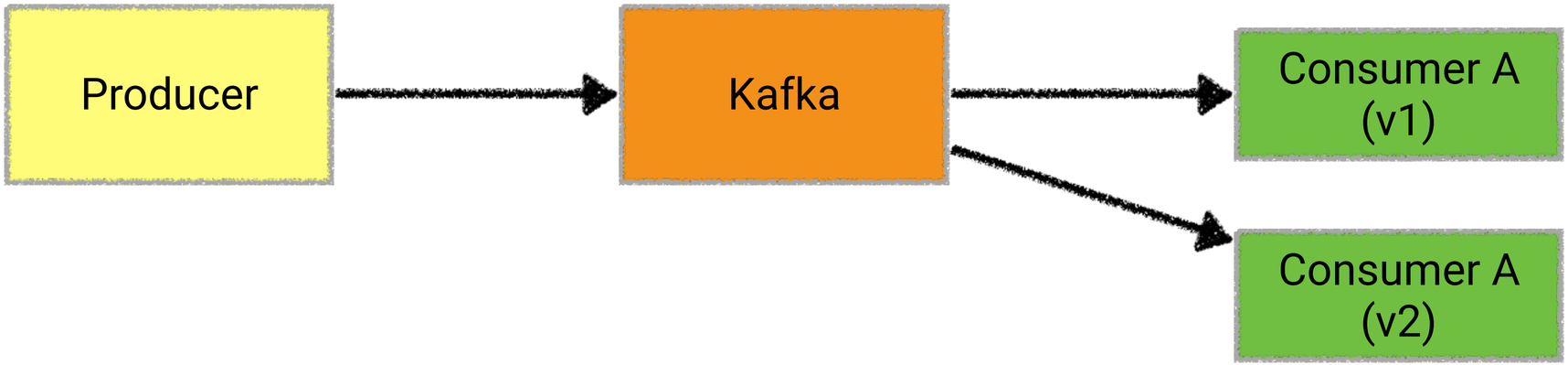


Varying Latency Requirements / Batch vs Stream

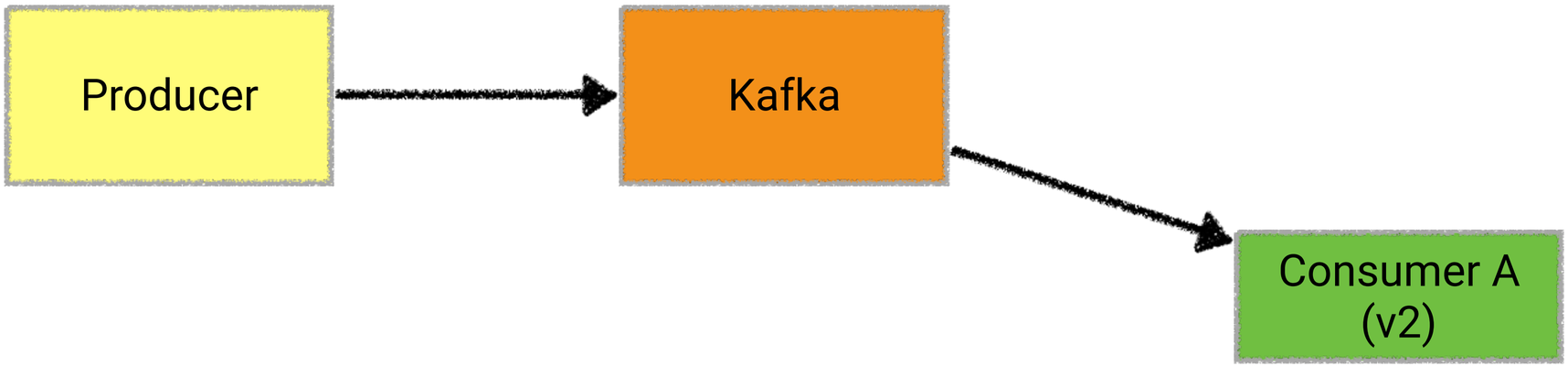


Technology & Code Changes

Technology & Code Changes

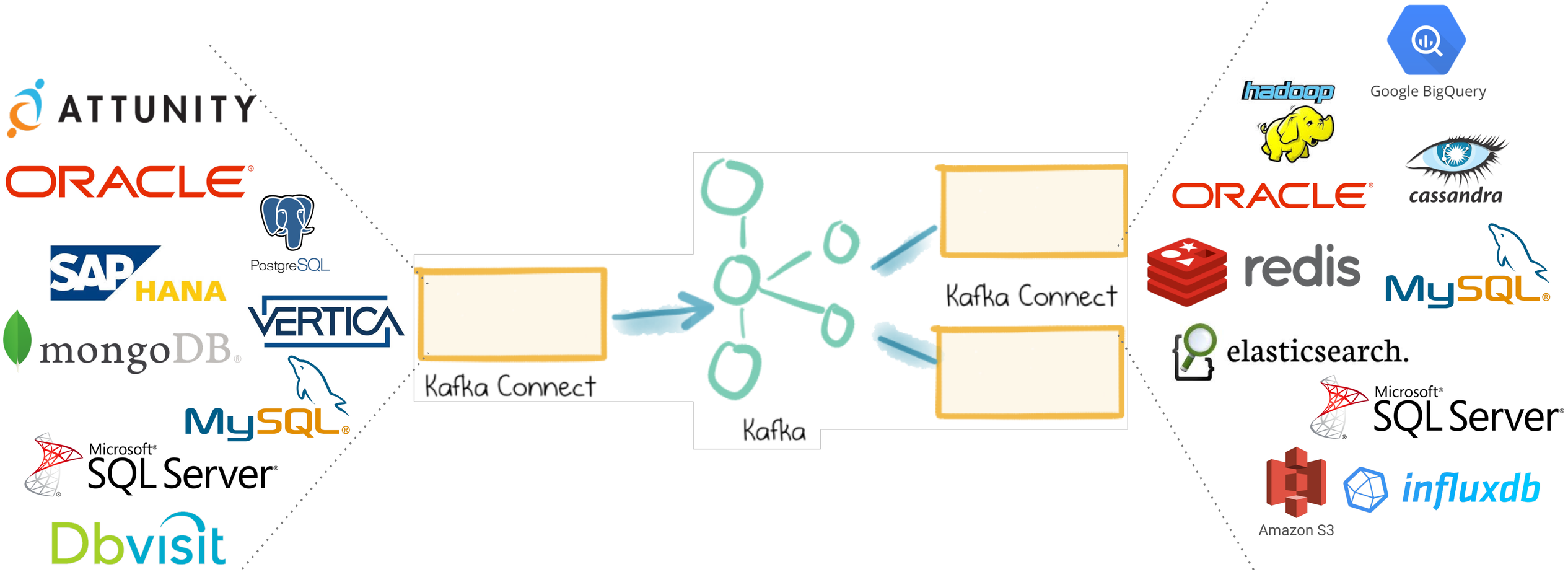


Technology & Code Changes



How do I get my data into Kafka?

Kafka Connect : Stream data in and out of Kafka

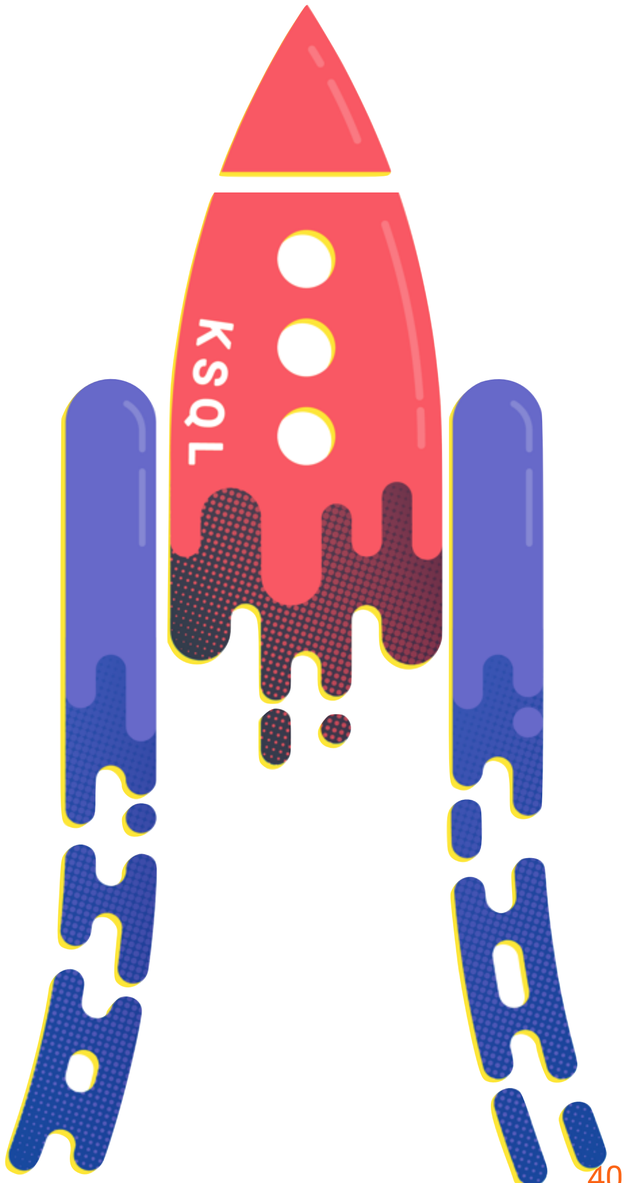


But I need to join...

aggregate...filter...

KSQL: a Streaming SQL Engine for Apache Kafka™ from Confluent

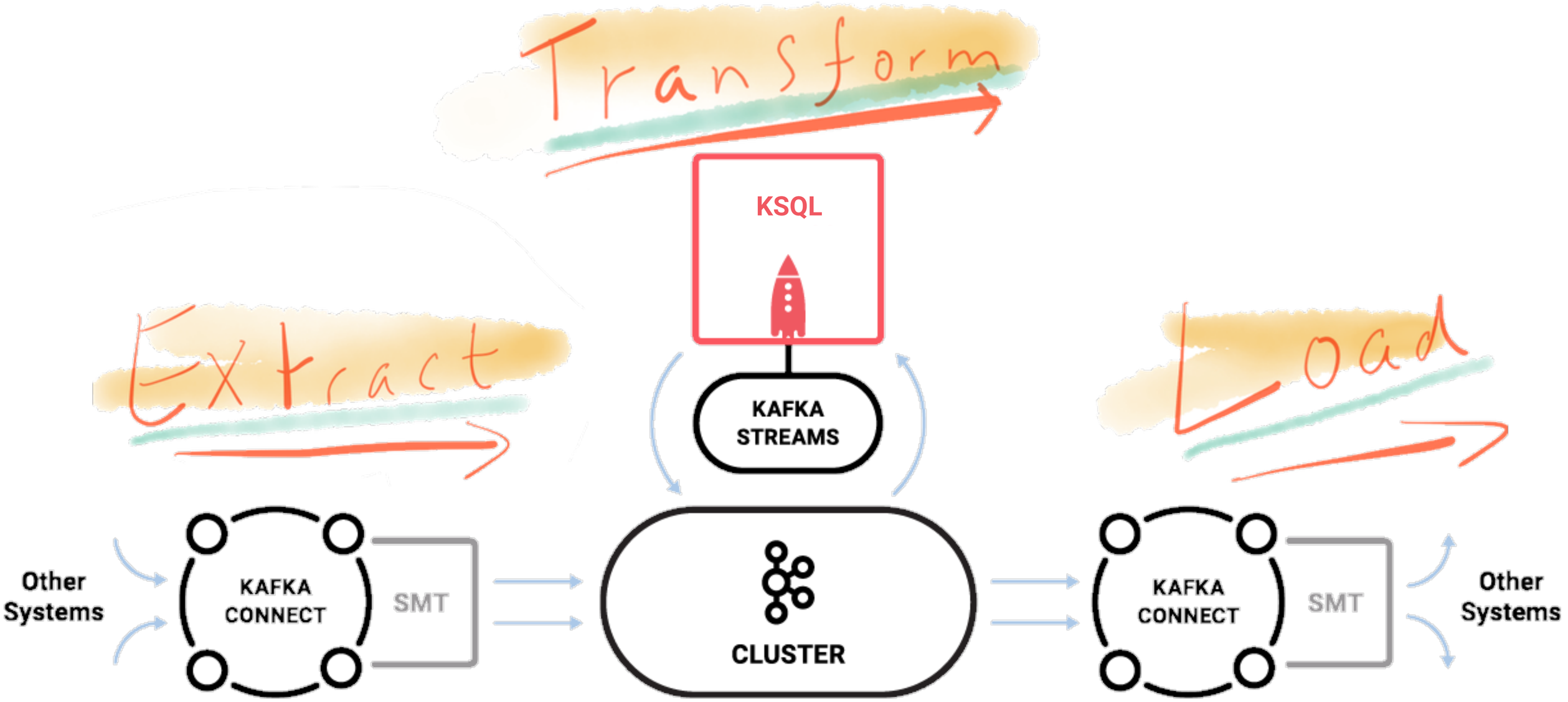
- Enables stream processing with zero coding required
- The simplest way to process streams of data in real-time
- Powered by Kafka: scalable, distributed, battle-tested
- All you need is Kafka–No complex deployments of bespoke systems for stream processing



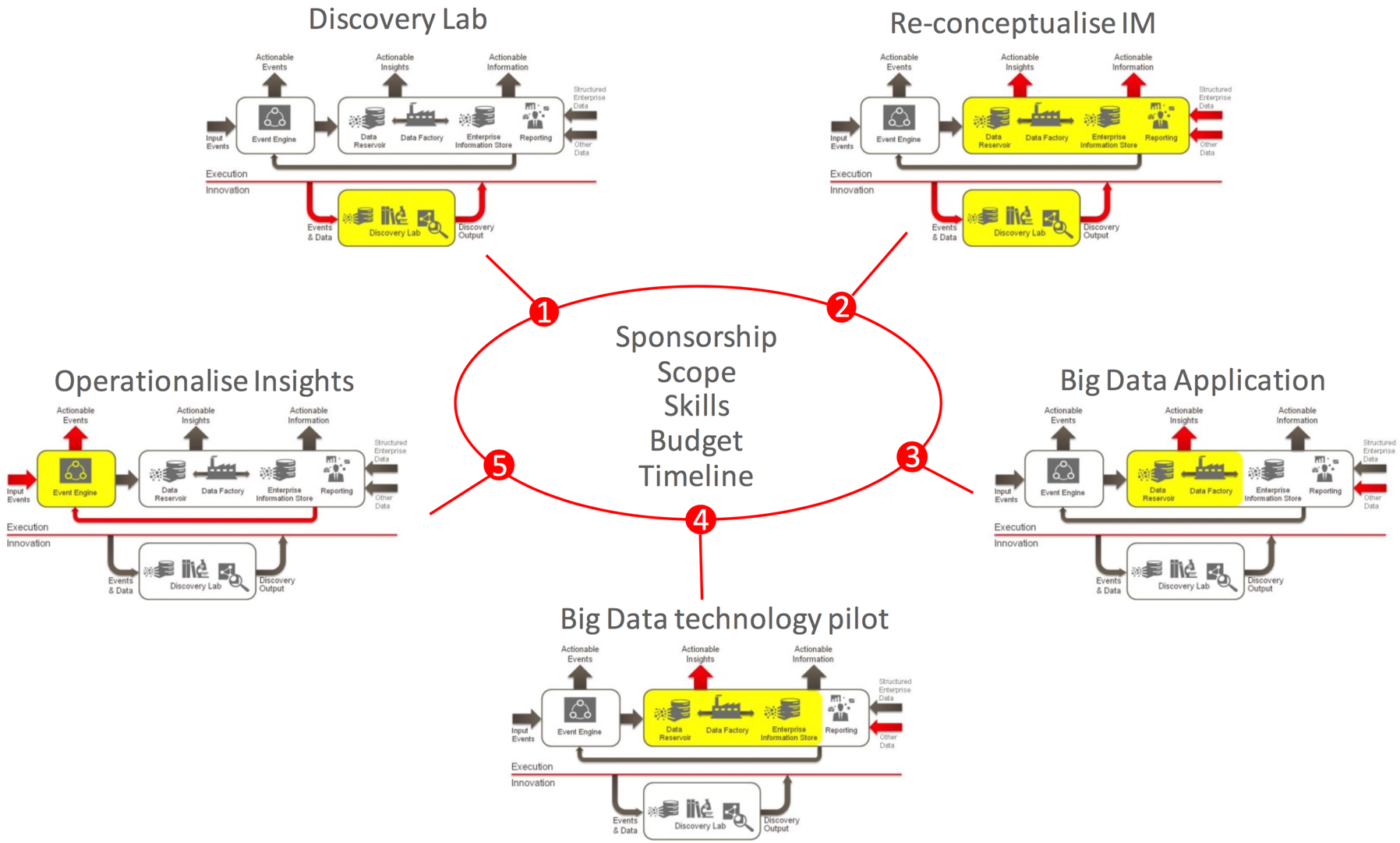
KSQL: the Simplest Way to Do Stream Processing

```
CREATE STREAM possible_fraud AS
SELECT card_number, count(*)
FROM authorization_attempts
WINDOW TUMBLING (SIZE 5 SECONDS)
GROUP BY card_number
HAVING count(*) > 3;
```

Streaming ETL, powered by Apache Kafka and Confluent Platform



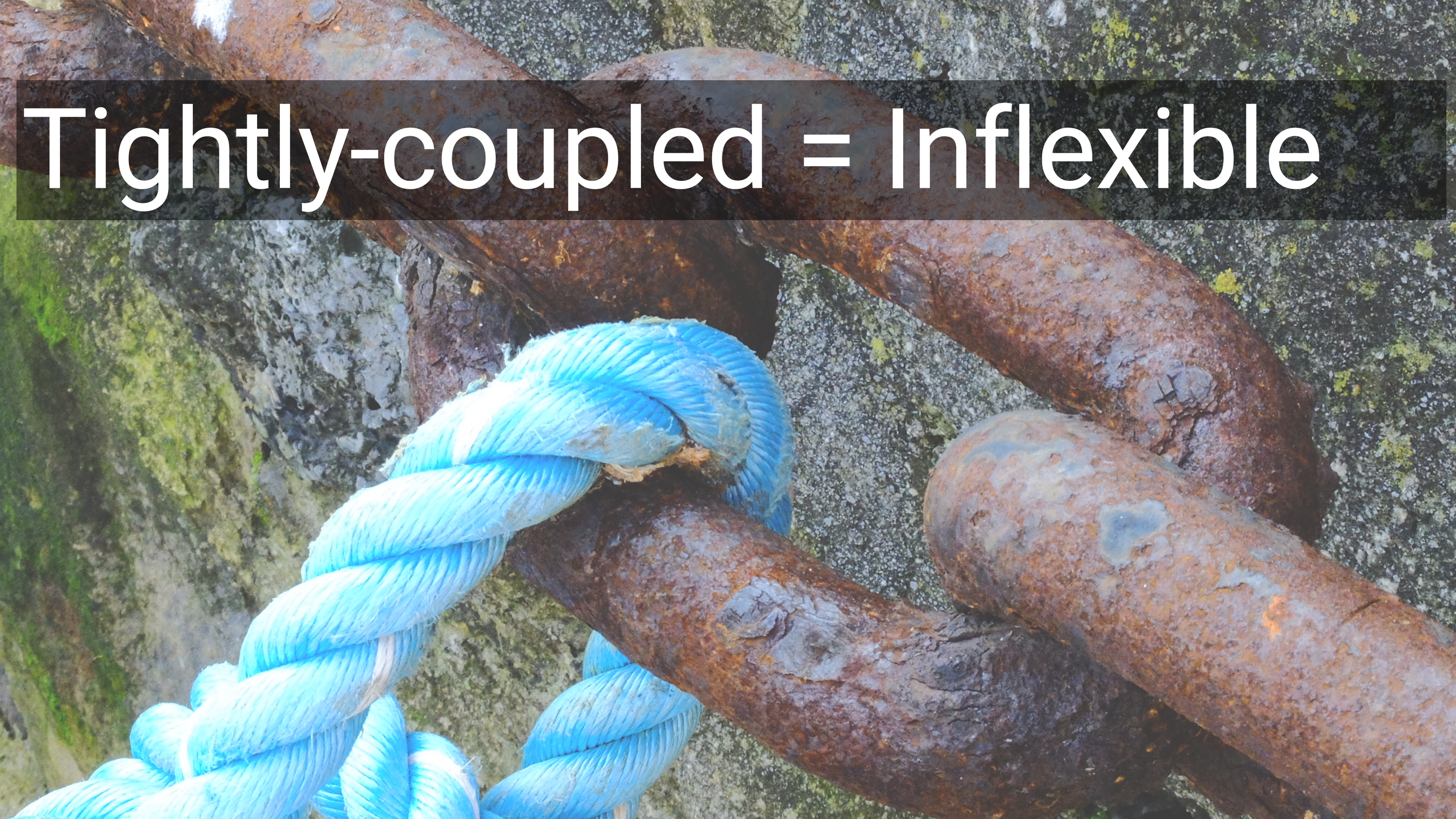
Reference Architecture - Implementation Patterns



http://www.oracle.com/technetwork/database/bigdata-appliance/overview/bigdatarefarchitecture-2297765.pdf

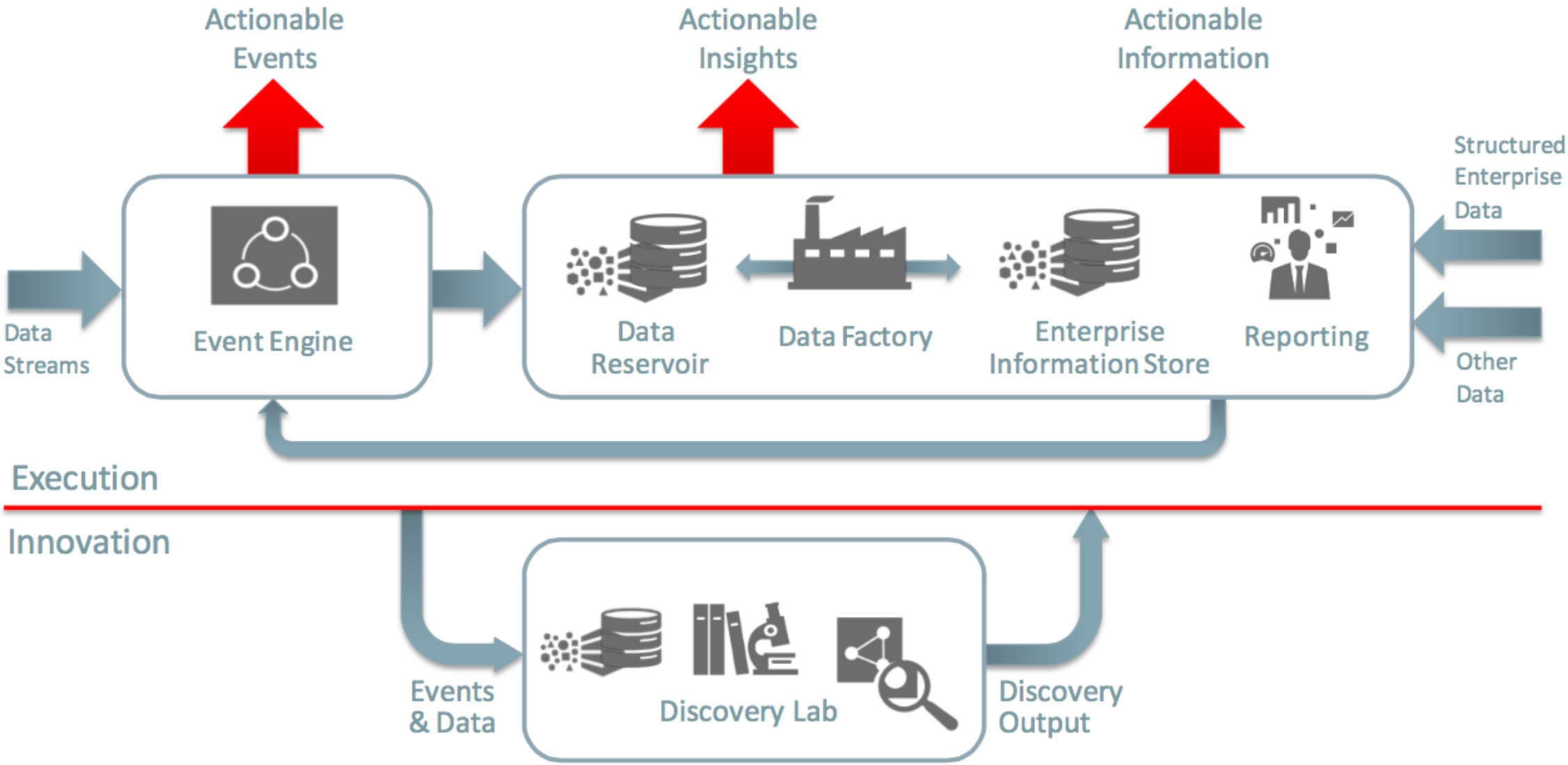
Building for the Future

- Enable flexibility & agility for:
  - Performance / scaling / resilience
  - Reducing latency / moving to stream processing
  - Increasing number of data sources
  - Connecting other (as yet unknown) consuming applications
  - Taking advantage of improved technologies (functionality, resilience, cost, scaling, performance)


Tightly-coupled = Inflexible

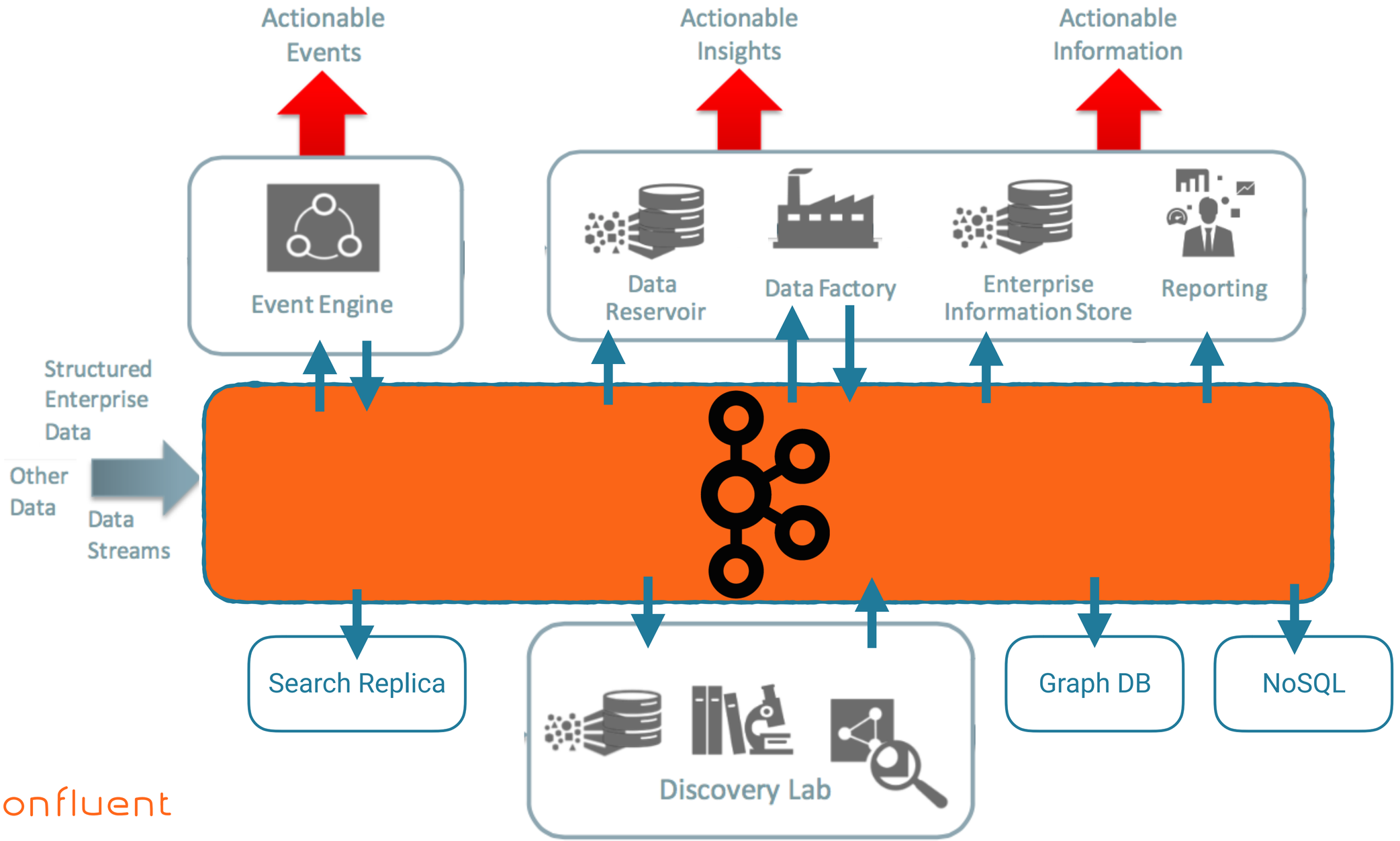
Apache Kafka is the solid foundation upon which you build *any* successful data platform

Conceptual Architecture

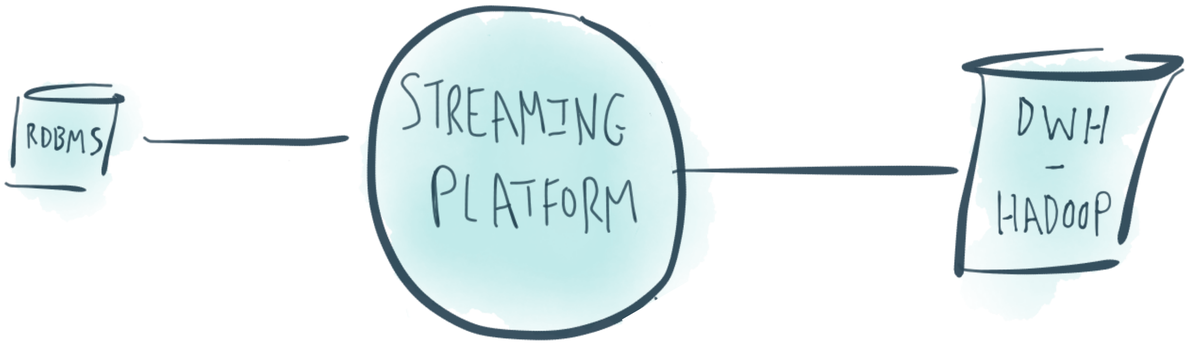


http://www.oracle.com/technetwork/database/bigdata-appliance/overview/bigdatarefarchitecture-2297765.pdf

Conceptual Architecture - with Kafka as the Backbone

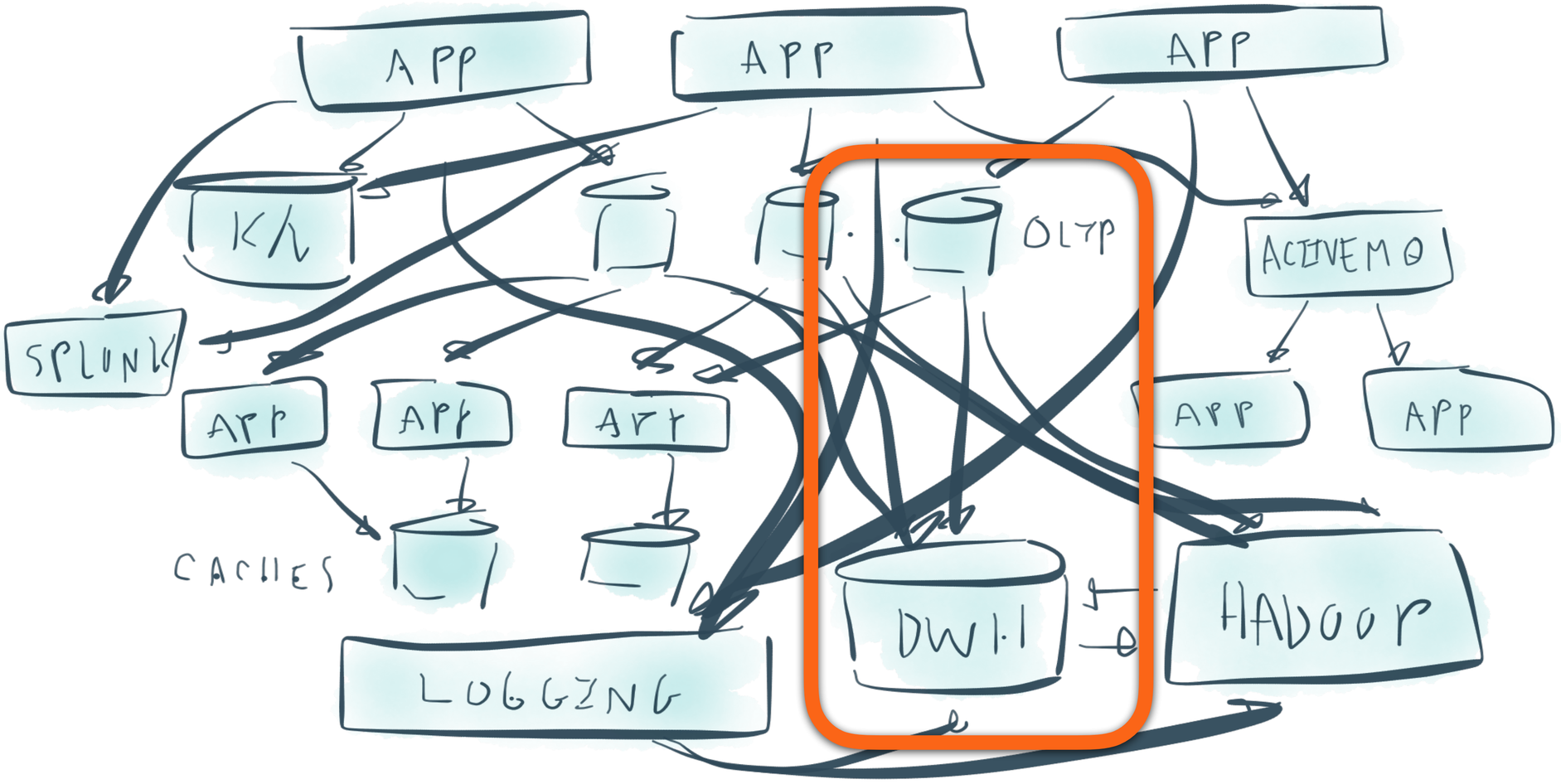


It's not just about Information Management and Analytics

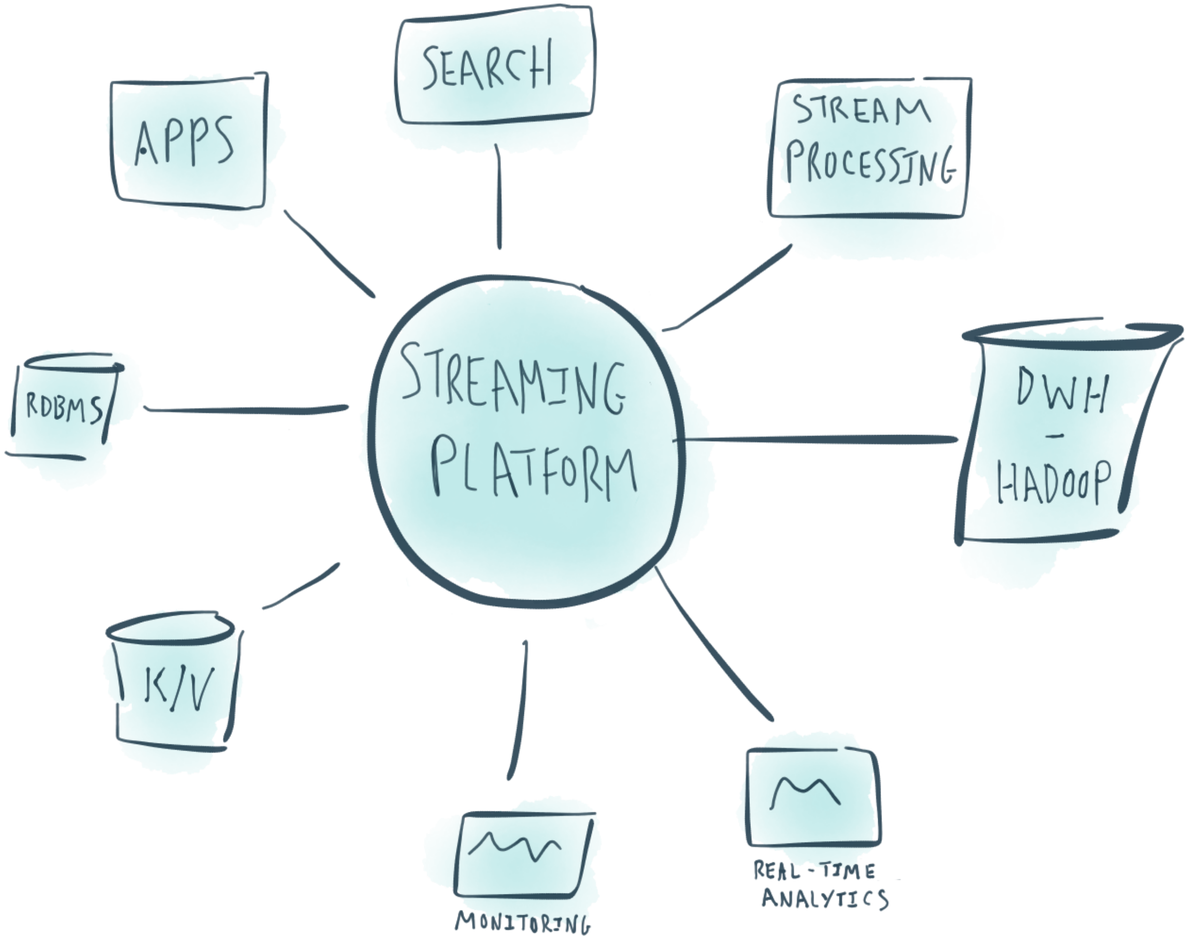


Best Tool -> Best Job

A GIANT MESS


APP
APP
APP
KAFKA
OLTP
ACTIVEMQ
SPLUNK
APP
APP
APP
APP
APP
CACHES
DW1.1
HADOOP
LOGGING

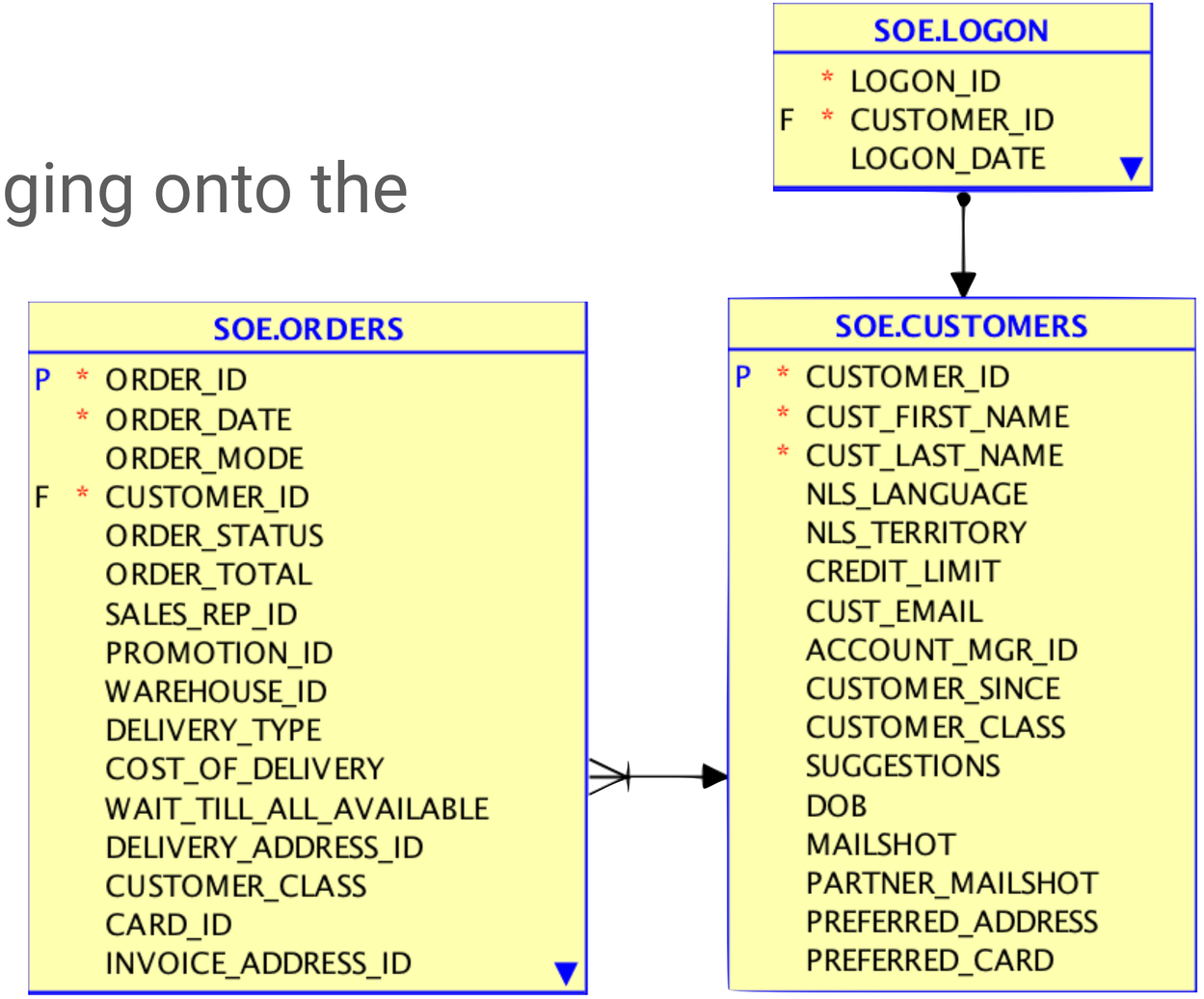
STREAMING PLATFORM



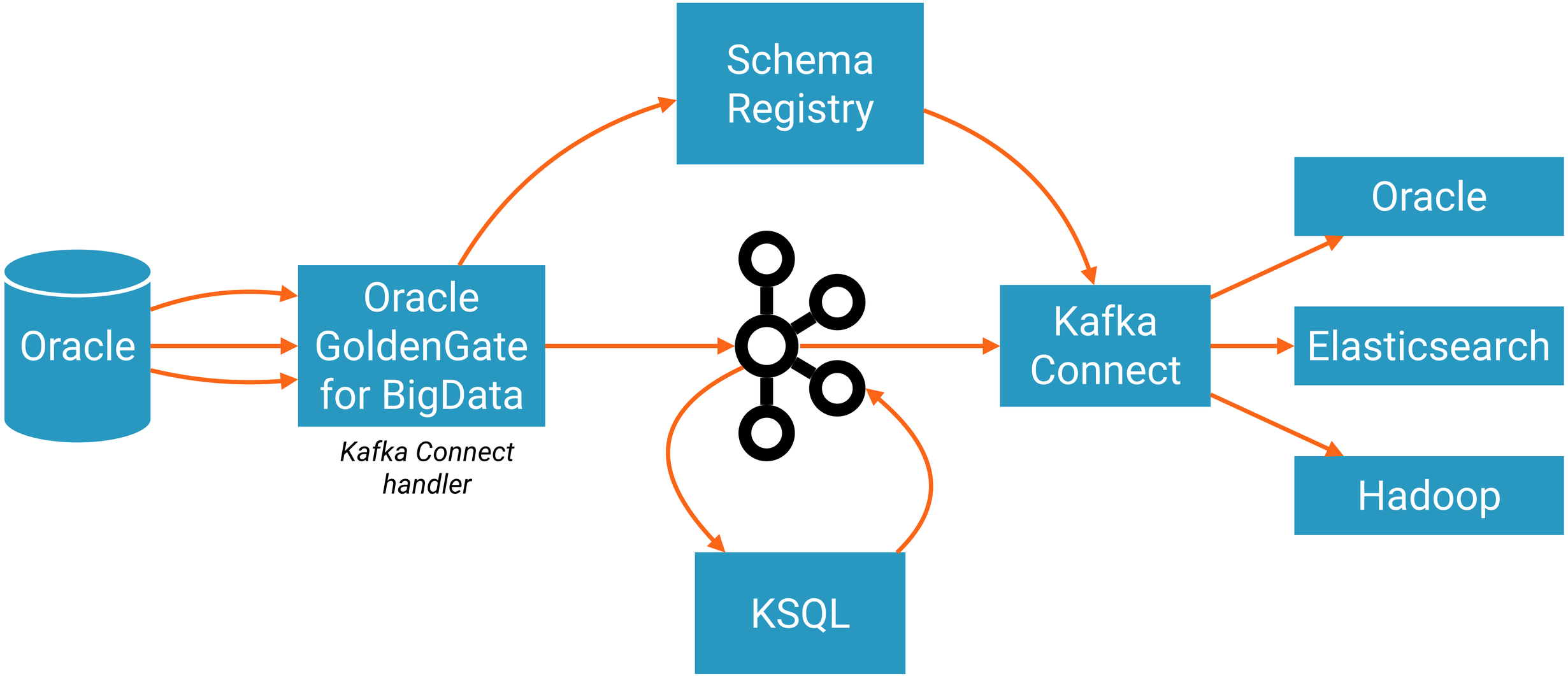
Putting it into Practice

Kafka for building Streaming Data Pipelines

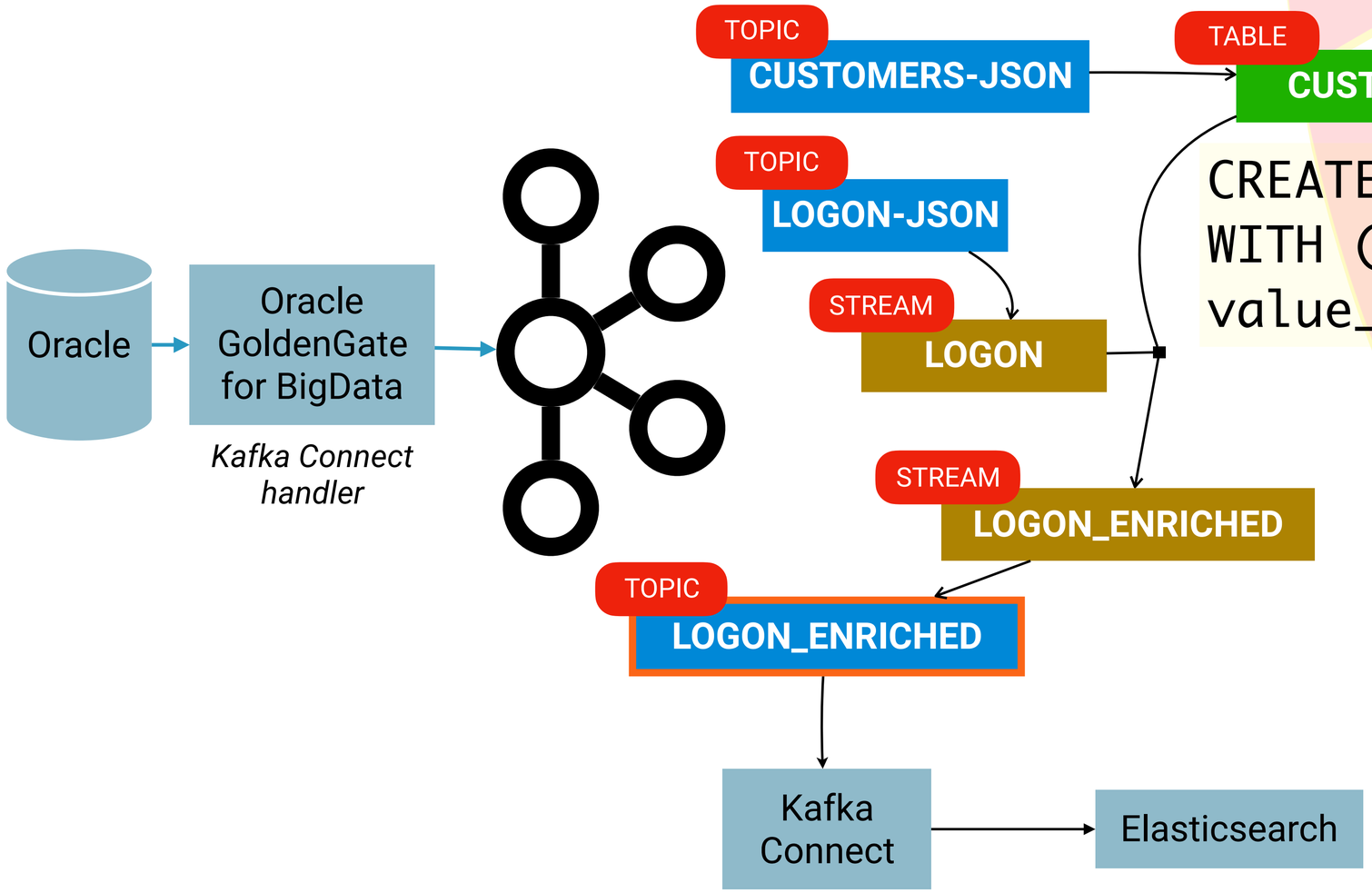
- Source data is online ordering system running on Oracle
- Requirement:
  - Realtime view of key customers logging onto the application
  - Realtime view of aggregated order counts and values
  - Long-term storage of data
  - Populate DW performance layer



Streaming ETL with Apache Kafka and Confluent Platform



```
CREATE TABLE CUSTOMERS (CUSTOMER_ID INT…)
WITH (kafka_topic='CUSTOMERS-JSON',
value_format='JSON');
```

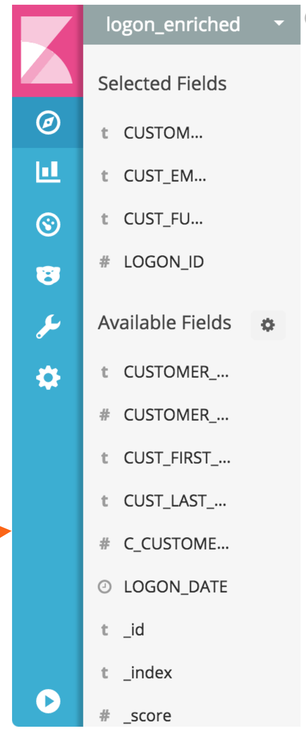


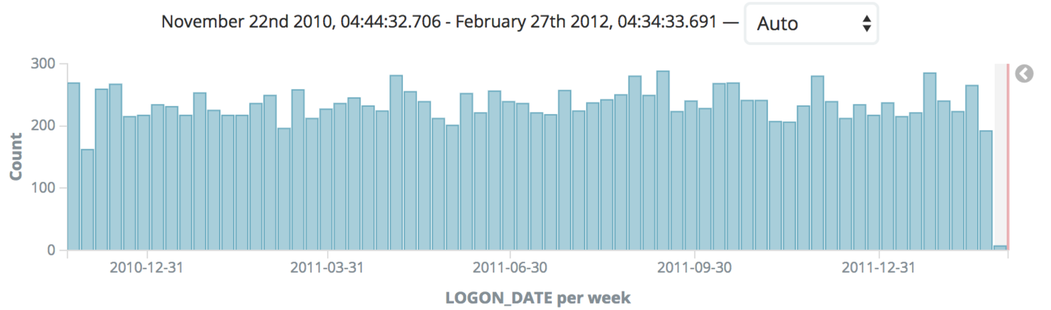

```
CREATE STREAM LOGON (LOGON_ID INT, …)
WITH (kafka_topic='LOGON-JSON',
value_format='JSON');
```

```
CREATE STREAM LOGON_ENRICHED AS
SELECT L.LOGON_ID, C.CUSTOMER_ID…
FROM LOGON L
LEFT OUTER JOIN CUSTOMERS C
ON L.CUSTOMER_ID = C.CUSTOMER_ID
```

Driving Realtime Analytics with Apache Kafka

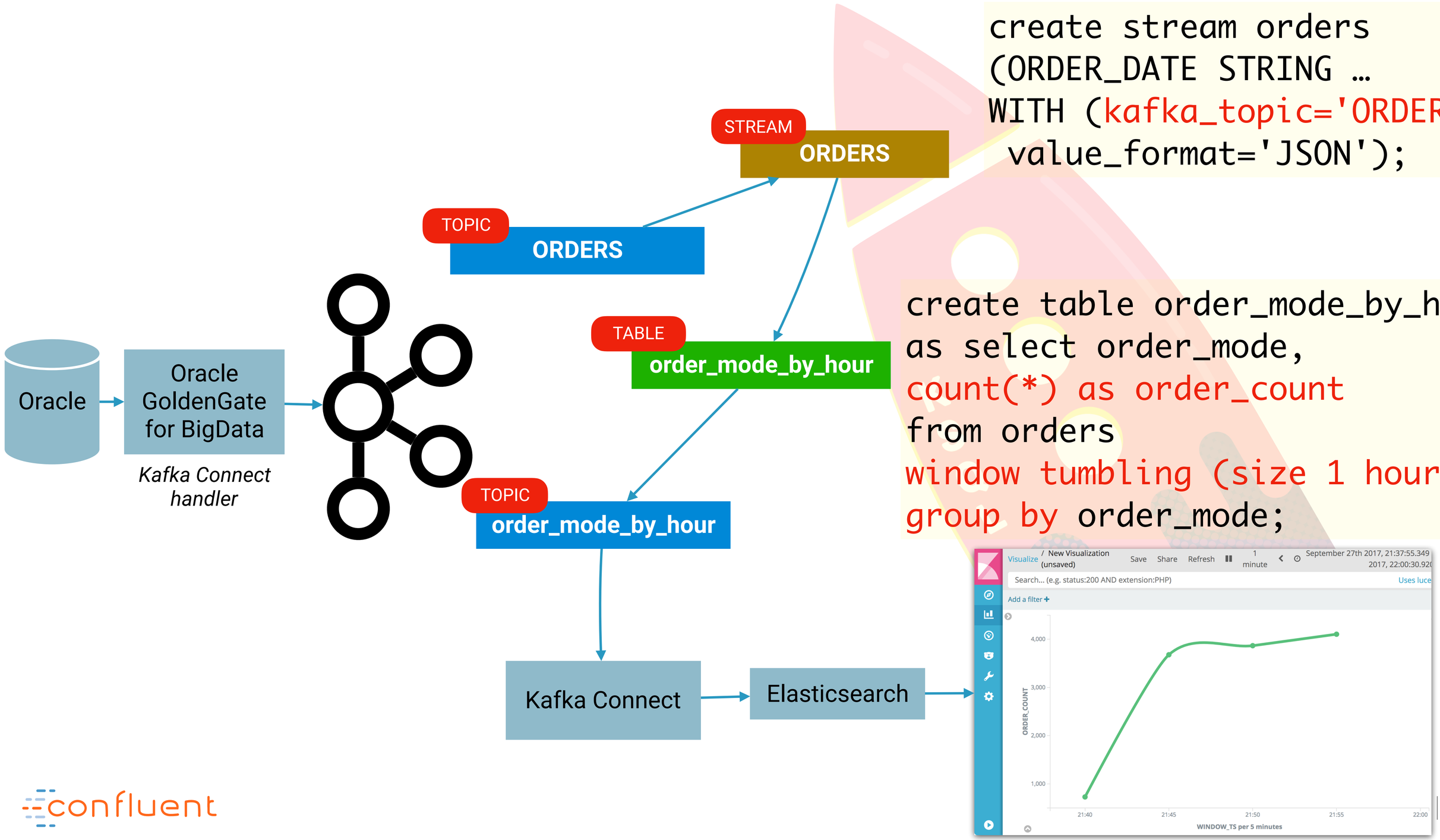
Time	NCR	UCD	BP	OP	PO	BO	SQ	WQ	WA	Users
17:36:29	0	0	0	0	0	0	0	0	0	[0/1]
17:36:31	0	1	0	0	0	0	0	0	0	[1/1]
17:36:33	9	7	47	33	2	7	0	1	1	[1/1]
17:36:35	25	16	155	78	7	24	0	2	3	[1/1]
17:36:37	49	25	276	137	14	39	1	2	4	[1/1]
17:36:39	75	35	395	195	16	55	5	3	4	[1/1]
17:36:41	108	45	506	262	25	76	8	4	5	[1/1]
17:36:43	137	56	625	334	30	87	9	5	6	[1/1]
17:36:45	165	66	785	405	37	111	12	5	7	[1/1]
17:36:47	191	83	947	487	42	134	16	5	8	[1/1]
17:36:49	219	97	1110	568	47	151	22	6	11	[1/1]



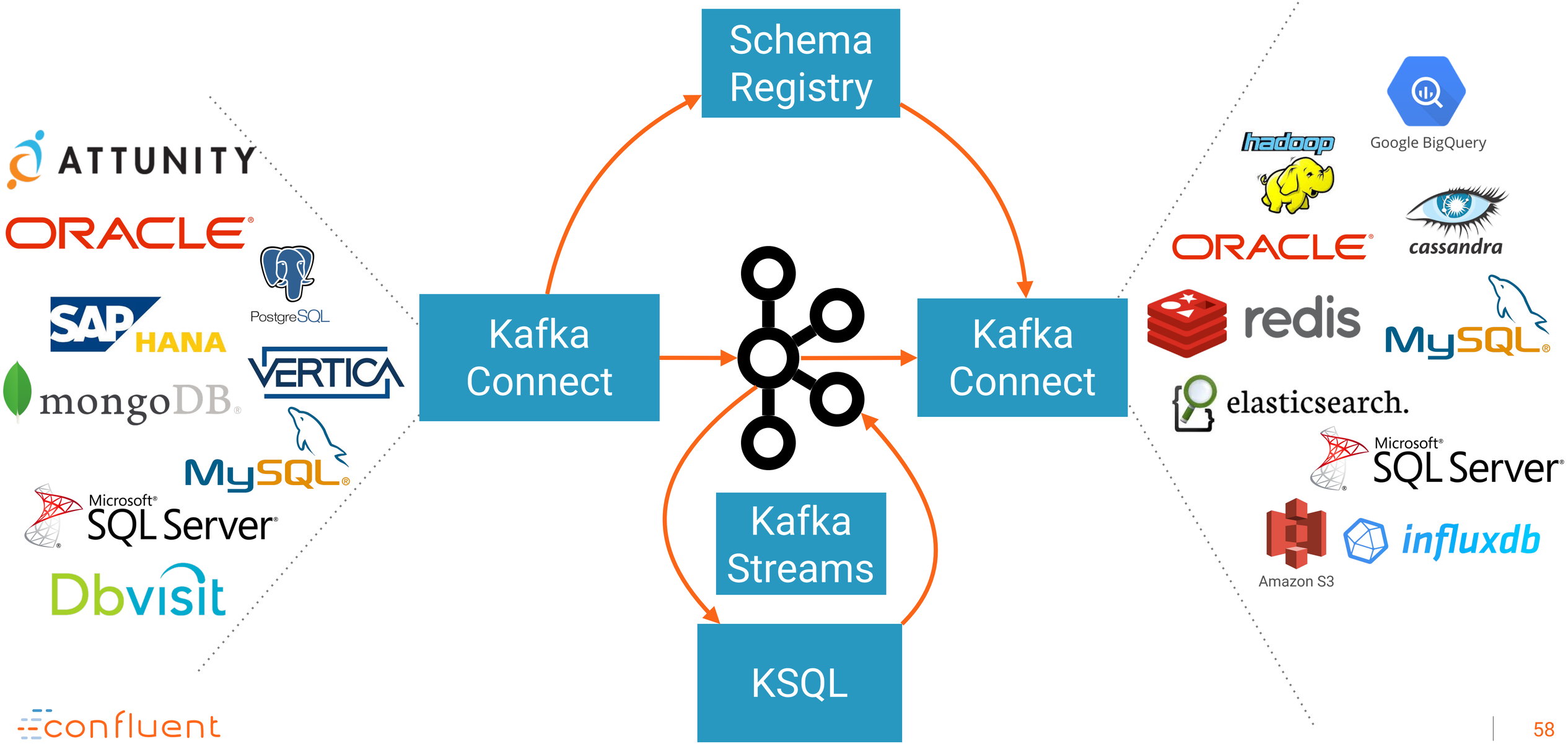


Time	LOGON_ID	CUST_EMAIL	CUSTOMER_CLASS	CUST_FULL_NAM
November 25th 2011, 00:57:35.000	40	byron.lucas@virgin.com	Regular	byron lucas
January 9th 2012, 12:40:49.000	65	julius.shelton@googlemail.com	Occasional	julius shelto
September 13th 2011, 10:43:59.000	72	alex.dixon@aol.com	Regular	alex dixon
August 26th 2011, 16:05:41.000	90	guy.weber@bellsouth.com	Prime	guy weber
December 25th 2010, 04:21:16.000	97	cesar.mckinney@aol.com	Regular	cesar mckinne

Events in the source system (Oracle) are streamed in realtime through Kafka, enriched via KSQL, and streamed out through Kafka Connect into Elasticsearch.


STREAM
ORDERS
TOPIC
ORDERS
TABLE
order_mode_by_hour
Oracle
Oracle GoldenGate for BigData
Kafka Connect handler
TOPIC
order_mode_by_hour
Kafka Connect
Elasticsearch
create stream orders
(ORDER_DATE STRING …
WITH (kafka_topic='ORDERS',
value_format='JSON');
create table order_mode_by_hour
as select order_mode,
count(*) as order_count
from orders
window tumbling (size 1 hour)
group by order_mode;
Visualize
New Visualization (unsaved)
Save
Share
Refresh
1 minute
Search... (e.g. status:200 AND extension:PHP)
Add a filter
4,000
3,000
2,000
1,000
ORDER_COUNT
21:40
21:45
21:50
21:55
22:00
WINDOW_TS per 5 minutes

Streaming ETL with Apache Kafka and Confluent Platform



500+ Technical Experts Helping Peers Globally

ORACLE ACE PROGRAM

ORACLE ACE Director

ORACLE ACE

ORACLE ACE Associate

---

3 Membership Tiers

- Oracle ACE Director
- Oracle ACE
- Oracle ACE Associate

bit.ly/OracleACEProgram

Connect:

oracle-ace_ww@oracle.com

Facebook.com/oracleaces

@oracleace

ORACLE Developer Community

Nominate yourself or someone you know: acenomination.oracle.com

ODTUG Kscope18

ORLANDO, FLORIDA • JUNE 10-14

Register Now

www.kscope18.odtug.com

WALT DISNEY WORLD SWAN

WALT DISNEY WORLD DOLPHIN

O R L A N D O

kafka summit
Kafka Summit 2018
Kafka Summit London: April 23-24, 2018
Kafka Summit San Francisco: October 16-17, 2018
www.kafka-summit.org
Presented by confluent

t: @rmoff e: robin@confluent.io
https://speakerdeck.com/rmoff/

Apache Kafka's Role in Implementing Oracle's Big Data Reference Architecture

SUN6259

Oracle OpenWorld, 1 Oct 2017

Robin Moffatt, Partner Technology Evangelist, EMEA

https://www.confluent.io/download/